"Stepping Stones to Street-wise Statistics"

Ken McKelvie : UoL Maths Club 27 March 2010

PART 1: **Conundrum**

From time to time stories of major decisions based on the misinterpretation, or disregard, of official data appear in our national media. Examples might be (i) unacceptable hospital-specific mortality rates attributable to mismanagement in the National Health Service or (ii) gross overspends on publicly-funded capital projects. If we assume that the education of the providers and the receivers of the data included mathematics to at least GCSE or A-Level, this assumption raises the questions "Why do the misinterpretations arise?" and "How might they be avoided?

PART 2: **Our Stepping Stones**

Data may be regarded as consisting of atoms of information. These may be quantitative or descriptive in character – or a little bit of both. Quantitative data are the substance of the study of statistics. Statistics has sometimes been described as "the art of asking the right questions and the science of getting the right answers". Statistical reasoning depends on clarity of thought and a great deal of common sense.

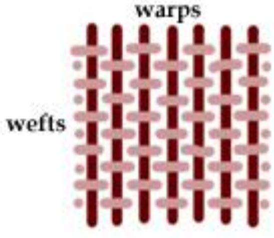
Maths-techniques feature strongly in the higher echelons of the subject. Today's presentation concentrates on the appreciation of the importance of common-sense reasoning – referred to here as "street-wise statistics". The maths-techniques elements will be left to another day.

Our aim is to introduce some of the language of statistics. The stepping stones are the use of pictures or summary-tables ("visualisation"), precision in the use of words ("verbalisation"), and plausible checking of presented outcomes ("validation"). Selected historically well-rehearsed examples of applications will highlight these steps "in use".

As you meet each of the examples in this presentation, see if you can identify which stepping stone is being used where.

PART 3: **Statistical reasoning – an informal view**

Regard, temporarily, "statistical reasoning" as being woven like a cloth. View (i) the warp, the threads length-wise, as being the Maths-techniques content including a range of probability reasoning and numerical computation algorithms and so on, and (ii) the weft, the threads woven cross-wise, as being the Data content – including collected quantitative evidence as number values and an awareness of the influence of all assumptions made within any analysis. A judgement as to an appropriate balance of inputs is required.



A simplistic view is that the strength of any statistical reasoning is weakened if too heavily projects the Maths-techniques without sufficient awareness being taken of the context of the available Data or, the other way round, if too little awareness of the underlying Maths-techniques leads on to an almost-blind overreliance on recipe ("book-work") analyses of the available Data.

PART 4: For awareness only – not developed any further here **: Human brain functions**

Human brain functions can be roughly classified as "detail/logic" oriented and as "big picture/feelings" oriented. Up to the time when brain-imaging technologies became available the former functions were attributed to the left half of the brain and the latter to the right half.
In terms of these classifications the study and practice of Statistics may be regarded as using both halves of the brain.
http://www.dailytelegraph.com.au/news/weird/the-right-brain-vs-left-brain/story-e6frev20-1111114577583

PART 5: **"What's the problem?" - Interpretation of words**

The interpretation of words matters, both in general and in Statistics in particular. The scope for misinterpretation and consequential misunderstanding is boundless.

The words "problem" and "solution" are both worth addressing.

"Problem" has two common interpretations. The first is as "a difficulty encountered which stops you getting from where you are to where you wish to be". The second is as "a specification of a task to be completed".

"Solution" also has two common interpretations. The first is as "a key to overcoming a difficulty". The second is as "a statement of the steps taken to get from your starting point to where you want to be".

PART 6: For awareness only: **On the lighter side**

(i) **The Two Ronnies' "fork handles" v "four candles" sketch**

The same or similar sounding words can have very different meanings in practice depending on the context of use. Communication between two people or two groups of people can fail if this is not recognised. Always seek an alternative interpretation and, if there is one, double-check as soon as practicable which is the intended one.

(ii) **Donald Rumsfeld's "unknown unknowns" quotation**

Donald Rumsfeld (b1932), an American politician and businessman, served as the 13th Secretary of Defense from 1975 to 1977 under President Gerald Ford, and as the 21st Secretary of Defense from 2001 to 2006 under President George W. Bush. One claim to fame was his observation that
"There are known knowns; there are things we know that we know.
There are known unknowns; that is to say there are things that, we now know we don't know.
But there are also unknown unknowns – there are things we do not know we don't know."
Think about it !
http://www.brainyquote.com/quotes/authors/d/donald_rumsfeld.html

http://www.youtube.com/watch?v=GiPe1OiKQuk

PART 7: For awareness only: **Selected noteworthy "statistical" events**

c0000: The "First Christmas" story. Saint Luke – New Testament. A census, ordered by Caesar Augustus, required Mary and Joseph to attend Bethlehem, to seek to ensure that everyone in the Roman Empire was paying their due taxes. (Historians ponder the precise dating)

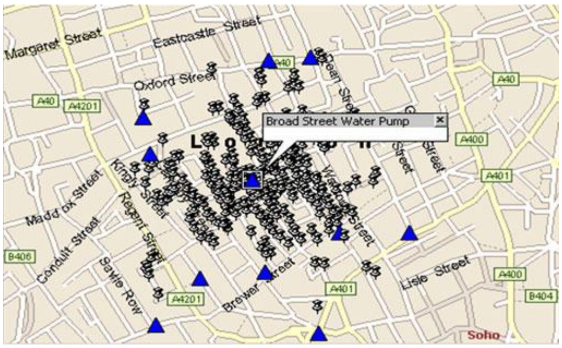
1086: The Domesday Book. William the Conqueror's England-wide survey of what or how much each landholder had in land and livestock, and what it was worth. Subsequently, over time, such measures of the State became regarded as "state-istics".

1791: Sir John Sinclair, a Scot, coined the wider use of the word "statistics" as applied to the study of numerical data arising in a more general range of contexts.

1834: Founding of the Statistical Society of the Statistical Society of London which, in 1884, became the Royal Statistical Society.

PART 8: **Three brief illustrations of Stepping Stones : leading to beneficial outcomes in practice**

(i) **John Snow (1813 – 1858) and the Broad Street Pump**
Cholera was originally believed to be an air-borne disease. There was a major outbreak in Soho in 1854. Over 120 deaths occurred in the first three days in September. By plotting places of death on a map Dr John Snow linked them to a particular water-pump located in Broad Street. Further investigation confirmed that cholera was water-borne and the need for the provision of an uncontaminated public water-supply was established.
http://www.makingthemodernworld.org.uk/learning_modules/geography/05.TU.01/?section=2



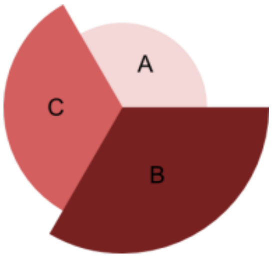
(ii) **Florence Nightingale (1820 – 1910)**
Florence Nightingale is well known for her nursing of the sick, and her pioneering reform of healthcare. Less well known is that she was also an accomplished statistician

In the Crimean War (1854-56), as common at the time, there were more deaths amongst soldiers from disease than as a result of battle. While Florence Nightingale was working at Scutari Hospital as the Superintendent of Nurses, the monthly rate of mortality in the first winter reached 40%. Eighteen months later, by improving hygiene, she was able to reduce the rate of mortality to 2% proved to her that much of the suffering of the army was unnecessary.

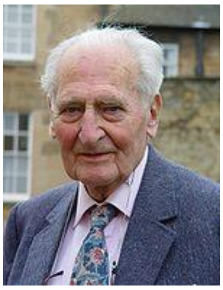
In 1858 she became the first female to be elected as a Fellow of the Royal Statistical Society.

Florence Nightingale made use of "coxcomb" representations (as exemplified below – and described in the given link) but, we are told, she did not refer to them as such. They are no longer in common use.

http://www.york.ac.uk/depts/maths/histstat/passionate_stat.htm
http://understandinguncertainty.org/node/215



(iii) **Richard Doll 1912 – 2005**
In 1951his team confirmed the link between smoking and the risk of lung cancer
http://news.bbc.co.uk/1/hi/health/3826939.stm



For other notable statisticians, see: http://www.bized.co.uk/timeweb/reference/statisticians.htm

PART 9: **Comparative** s**ubjective estimation or risk**

Participants in groups were invited to suggest and then rank in descending order of importance six requirements that they felt should be fulfilled for a journey by rail from Liverpool Lime Street to London Euston to be considered "fit for purpose".

A typical listing was
- travel safely
- get to London
- on-time
- in comfort
- with "good coffee"

Post-hoc it was noted that (i) the probability of a requirement being unfulfilled increased inversely to the cost of unfulfilment (ii) interest in the next one down the list is only triggered if its preceding requirements are fulfilled.
The exercise was intended to demonstrate the general commonality ranking of risks amongst individuals having similar prior experiences.

PART 10: **"Wimbledon"** : **Model validation and interpretation**

Data were gleaned from: Wimbledon Men's Finals
http://all-about-tennis.com/menswimbledonfinals.html

The following table shows, for the preceding forty years, those "Years" in which the Men's Final ran to 3 sets, 4 sets, and 5 sets.
Totals and proportions are computed

Match won in:	3 sets	4 sets	5 sets
Years	1973	71	70
	74	75	72
	76	81	77
	78	85	79
	83	88	80
	84	93	82
	86	85	90
	87	00	92
	89	04	98
	91	06	01
	94		07
	96		08
	97		09
	99		
	02		
	03		
	05		
Total No of Years	17	10	13
Observed Proportions	0.425	0.250	0.325

Next, the possible outcomes for players "Head" and "Tail" were listed for the matches won by "Head" and counted. Similarly for the matches won by "Tail". The counts for "Head" and "Tail" were combined.

Working with the assumption that finalists were of matched standards the corresponding probabilities were then calculated:

Match won in:	3 sets	4 sets	5 sets
Calc Probabilities	0.250	0.375	0.375

(These calculations were repeated by use of "formulae")

Discussion point: We note that in fact a greater proportion of matches were completed in three sets than would have been anticipated for players of matched standards. It appears that in unmatched opponents are more likely than matched. Detailed inspection of the source results might indicate that particular individuals have "waxed" and "waned" over their years of dominance.

PART 11: **Simpson's Paradox**

From an early age we learn the lesson that percentages alone are effectively meaningless. We need to know for each percentage the total to which it refers.

It is through overlooking the applicability of this cautionary lesson that you can trick or deceive your otherwise rational reasoning to lead you to potentially nonsense or otherwise paradoxical outcomes. In particular it is too easy to let slip, or not appreciate, the effects of any assumptions that you might be implicitly making.

Whilst well-known earlier, the incidence of the "paradox" bearing his name was spotlighted first in 1951 by Edward H. Simpson, a British statistician born 1922.

Hypothetical data showing outcomes of A-level Maths exams in a small town for students from the local School Sixth Form and from a competing Further Education College

		Pass	Fail	Total	Pass Rate %
Girls	School Sixth	7	3	10	70
	Further Ed	18	12	30	60
Boys	School Sixth	9	21	30	30
	Further Ed	2	8	10	20

The interpretation might be that for the girls the School Sixth Form produces a better pass rate than the Further Education College. A similar outcome holds for the boys.

The conclusion might be that overall the results for the School Sixth Form are better than those for the Further Education College.

HOWEVER, combining the results for the girls and boys in the School Sixth Form and similarly in the Further Education College the table becomes as follows.

		Pass	Fail	Total	Pass Rate %
Girls + Boys	School Sixth	16	24	40	40
Girls + Boys	Further Ed	20	20	40	50

If you are a prospective student, where would you elect to attend?
If you were a local decision maker required to retain provision for Maths students in one institution and withdraw it from the other, which would you choose to retain.

The competing outcomes are an example of Simpson's Paradox.

Key Point: In the first analysis we assume that we allow that the Maths abilities of girls and of boys might be different. In the second analysis we assume that there is no difference.

Think about it !

Further discussion and examples are available via the link :

http://www.cawtech.freeserve.co.uk/simpsons.2.html

PART 12: **"Statistics" Reading List**

Here are extracts from a 'universal' list that I have used for years. I have not checked recently but it is possible, maybe likely, that some of those texts listed are now out of print. They should however still be accessible via libraries.

Introductory

- Recommended
  "easy read" P. Sprent "Taking Risks" Penguin
  "read" C. Chatfield "Statistics for Technology" Chapman Hall
  "reference" G.M. Clarke & D. Cooke "A Basic Course in Statistics" Edward Arnold – in part aimed at A-level

- Other books - "non mathematical reads"
  A.S.C. Ehrenberg "A Primer in Data Reduction" Wiley – includes guidance on data display and report presentation
  P.G. Hoel "Elementary Statistics" Wiley
  D. Huff "How to lie with Statistics" Pelican

History of development of "Statistics"

The chronology and background of the development of the subject:.

- Highly recommended - if you are able to access a copy.
  J. Leroy Folks "Ideas of Statistics" Wiley

- The following undergraduate text is written to serve purposes above and beyond A-level requirements. It does however include accessible sections covering the chronology and background of the development of the subject.
  F. Daly, D.J. Hand, M.C. Jones, A.D. Lunn, K.J. McConway "Elements of Statistics" Open University Addison Wesley

An example of a context: "Will Pickles of Wensleydale" (1885-1969) by John Pemberton.

A powerful (and charming) contribution to the relevance of systematic data collection in a real-life (and death) context can be found in the very readable biography of "Will Pickles of Wensleydale" by John Pemberton. Will Pickles (1885-1969) was a GP in the 1920's in, you guessed, Wensleydale in Yorkshire. This was in the days before motor travel became commonplace. He recognised the value in recording contacts in the dale as the basis of the understanding of the spread of "epidemic" disease, particularly amongst children, and not only is his biography but also his data sets are still available in print

I recommend that you put the biography on your "sometime" reading list. The original edition, or the more recent edition republished by the Royal College of General Practitioners, should be available via local public libraries.

[Presentation ends]